16/PRTS

09/890168
Rec'd PCT/PTO 26 JUL 2001

JA998222

SPECIFICATION

METHOD AND DEVICE FOR DESCRIBING VIDEO CONTENTS

Field of the Invention

The present invention relates to a method and an apparatus for describing contents of motion picture for the sake of efficient search based on contents of motion picture of a computer system capable of processing motion picture, in particular, video, DVD and so on.

Background art

While motion picture such as video data is easy for a human being to understand, there was a difficulty for a computer to manage it. Namely, it is difficult to find the meaning that the contents have from an original video data itself, and it has not been possible to date to accurately represent the meaning of video data even with advanced image processing technology.

While management of motion picture such as video data with a computer was conventionally based on annotations made by an administrator in advance, the method of using annotations lacks consistency in annotations among administrators, and besides, complicated processing for video data in large quantities is a big problem from now on.

As a candidate for the solution, it is considered promising to describe metacontents from video data with an intermediate result of image processing and knowledge registrable in advance that the contents have.

However, though it is possible to use plural image features for specific contents so as to describe a description method or a search engine for specific search or management, general versatility is lost and it does not contribute to proliferation of video search.

Therefore, a description method with general versatility taking advantage of image features is desired for description of video data, and the activities for standardization were started by ISO (International Organization for Standardization) as MPEG-7.

MPEG (Moving Picture Experts Group) is an organization promoting standardization of encoding method for storing color motion picture, and MPEG-1, MPEG-2 and MPEG-4 have been standardized so far.

Since MPEG-7 has no rule for a method of image processing and is beyond the scope of any standards, not only automatic processing but also manual data input is allowed.

However, it will only make data input more complicated to demand meaning of originally unextractable scene from video data or registration of data difficult to detect from video data. So far, there have been many examples of representing frame sequence of video as structured. For instance, Abe's method ("Method for Searching Moving Picture with Change of State over Time as Key", Abe, Sotomura, Shingakuron, pp.512-519, 1992 (conventional example 1)) describes a dynamic change of state so that the time intervals to be searched in video search may not be fixed.

In Abe's method (conventional example 1), however, since information of state description covers the entire frames, a drawback is that search time is in proportion to length of

the video used for search. Also, since an object is represented by the center of gravity in an image, it is substantially different from the present invention taking advantage of changes in an object's shape.

While the method described in conventional example 2, ("An Automatic Video Parser for TV Soccer Games," Y.Gong, C.H-Chuan, L.T.Sin, ACCV '95, pp.509-513, Nov., 1995) is trying to use information of positions and movement of the players, the positions are classified code of positions where the field is roughly divided into nine, and movement is of a very short period (several frames) so that event extraction is performed regarding classified code of positions and motion vector in a short period as events.

In conventional example 2, however, a drawback is that an event to be extracted and description are inseparable and besides, extractable events become a very limited set.

The method described in conventional example 3, ("Integrated Image and Speech Analysis for Content Based Video Indexing," Y-L.Chang, W.Zeng, I.Kamel, R.Alonso, ICMCS '96, pp.306-313, 1996) adopts a limited approach of tracking positions of the

ball and goal posts on the screen and considering only their positional relationship so as to extract time intervals of close distance as exciting scenes.

The method described in conventional example 4, ("Analysis and Presentation of Soccer Highlights from Video," D.Yow, B.L.Yeo, M.Yeung, B.Liu, ACCV '95, pp.499-502, 1995) performs shot extraction covering American football, and identifies events such as touchdown by a key word in each shot by speech recognition and line pattern extraction in the screen using image processing.

However, neither conventional example 3 nor 4 has a concept such as a player and his movement.

On the other hand, while there is also conventional example 5, ("A Suggestion of a Method for Image Search by Inquiries Using Words of Movement," Miyamori, Kasuya, Tominaga, Image media processing symposium '96, I-8, 13, 1996) as a method for representation which cuts an object from video and is based on lifetime and an object position, it has neither a concept of reference plane nor general versatility.

In addition, conventional example 6, ("A Suggestion of a

Method for Image Contents Search Using Description of Short Time Movement in a Scene," Miyamori, Maeda, Echigo, Nakano, Iisaku, MIRU-98, I-75, 1998) also describes an object with description of short time movement as a unit, but it lacks expandability since it does not simultaneously adopt description which represents a spatio-temporal trajectory and is a representation method dependent on specific contents.

Conventional example 7, ("On the Simultaneous Interpretation of Real World Image Sequences and Their Natural Language Description: The System SOCCER," E.Andre, G.Herzog, T.Rist, Proc. 8th ECAI, pp.449-454, 1988) is a system with scene description and interaction among objects as its meta data. However, the purpose of the system of conventional example 7 is conversion of medium from image to speech, namely a system for automatically generating a narration, so it does not store the created meta data and unlike the present invention, it does not have a data structure suitable for search of contents.

Conventional example 8, ("Automatic Classification of Tennis

Video for High-level Content-based Retrieval," G.Sudhir, J.C.M.Lee, A.K.Jain, Proc. CAIVD-98, pp.81-90, 1997) covers tennis matches, so description of interaction among objects is limited to simple movement and position information.

The present invention is limited, in terms of its descriptive contents, to processing results based on "feature colors," "texture," "shape" and "movement."

In video, a subject of attention is different depending on its contents. Therefore, it is necessary to predefine a subject object depending on the contents.

An object defined here consists of a lump region appearing in image, and it is possible to extract its color, texture, shape and movement.

It is a property of this object region that can be extracted from video, and it is difficult to give meaning to its contents.

Accordingly, a technique of description based on relationship between a single object and plural objects is proposed, and knowledge dependent on contents registrable in advance and description of objects are associated, and thus,

search based on an object of a meaningful scene in video becomes possible.

Since description of the entire frames of video data will result in storing of redundant information in large quantities, description for efficiently representing video contents with a little data volume is important.

The present invention is a proposal of a description method effective for interpretation based on video contents. The description method of the present invention is effective not only for search of an object or a scene but also for applications such as reuse of an object and summarization of contents.

An object of the present invention is to provide a description method for efficiently representing contents of motion picture with a little data volume.

Another object of the present invention is to propose a description method effective for interpretation based on contents of motion picture.

A further object of the present invention is to provide a description method not only for search of an object or a

scene but also capable of applications such as reuse of an object and summarization of contents.

Summary of the Invention

The present invention proposes a method for description in terms of relationship between a single object and plural objects based on data extractable as an image feature.

Namely, the present invention (1) represents a trajectory of how each object has moved over time by using reference plane representing position information of each object, (2) sets a description unit based on a type of action of each object by using changes in shape of each object, (3) has actions of each object represented as each behavioral section, and (4) comprises a description facility capable of reading and interpreting definition of an object dependent on video contents, definition of classes of actions, and definition of interpretation of a scene by interaction of plural objects.

Brief Description of the Drawings

Fig. 1 is a diagram showing cutting of video data on a computer to which the present invention is applied.

Fig. 2 is a diagram showing that position information and actions are assigned to a behavioral section of each object to which the present invention is applied.

Fig. 3 is a diagram showing a trajectory on reference plane of each object to which the present invention is applied.

Fig. 4 is a diagram showing an outline of a processing flow of video data on a computer to which the present invention is applied along with its main components.

Fig. 5 is a diagram showing a search screen of the present invention.

Fig. 6 is a diagram showing a search result of the present invention.

Preferred Embodiments

As main features of an object in image, there are "position" of an object, its "moving speed" and "trajectory."

However, in spite of conventional use of "position" information, position in description of video is mostly mere

implicit use of an image plane, and there is no method for using a different ground from the image plane depending on the contents.

In addition, an image is originally a three-dimensional scene projected on a two-dimensional plane, and there are cases where it is more convenient to take an object position not on an image plane but on a plane on which the object exists.

In general, sports images just fall under the cases where it is more convenient to take an object position not on an image plane but on a plane on which the object exists.

While a plane equivalent as world coordinates system is generally used for description of an object of the real world, it is different from video in the subject of description, purpose and in that it always takes the ground for a part of the real world.

On the other hand, since an image lacks information of a depth direction, it may be better to project consecutive frames of video on a common image plane.

From the above consideration, it is necessary to predefine

for each of the contents to be searched the ground for determining an object position which is called reference plane and described as follows by using "zone description" and a "camera model."

This description method allows effective description for the contents to which a geometrical relationship among objects is important.

Descriptions of reference plane, zone description and camera spec. are shown in attached Table 1. However, the description of camera spec. is not essential to the present invention but optional.

Next, representation method for each object is explained. An object moves on reference plane and performs meaningful behavior/actions in it.

For the contents in which behavior of an object may be a major factor in hitting the search, specifically the following description method is very effective.

A representation unit of an object is decomposed on the basis of behavior, and initiating and terminating frames representing its behavioral section and a trajectory between

them are described so as to reproduce an object position in an arbitrary frame.

While nothing represented changes in shape of each object in the background art, the present invention allows meaning of an object of a description unit to be preserved by giving meaning to changes in shape as behavior of an object due to such changes in shape.

This description is represented as Action (or described as Act) shown in attached Table 2.

An example of description using this Action is as follows.

Act("Kick", 10(frame), 240(frame), 3, 3, (120, 180, 0) (150, 195, 180) (180, 223, 230))

By this description of Action, a representation of "An object (who) did the behavior indicated by action ID (what) on the space represented by a trajectory (where) in a time interval (when)" becomes possible.

Here, collinear approximation is used as a method for representing a trajectory. Data of a trajectory is

collinearly approximated, and the number of nodes, approximated coordinates on reference plane of each node and the time in Action are described. Consequently, if a certain time is specified, coordinate value of an object at the time can be uniquely determined.

Fig. 1 is a diagram conceptually representing description based on reference plane and a behavioral section of an object and description and data flow based on plural objects in soccer contents.

First, a video object based on the region is cut from video sequence, and lifetime of the object is acquired by tracing each object in the direction of time (131). Next, actions of the object are classified based on silhouette of the object which represents changes in shape of each object. Description of objects is performed for each of these behavioral sections, which is Action description (132).

At this time, spatial movement of the object is represented as its trajectory by using reference plane. Objects present in the entire frames of video have Action description independently, and IAction (described later) is defined from

plural objects.

For simplicity, a trajectory in Act is described here limiting it only to that of two points. In addition, a trajectory of a ball is used as to spatial representation of IAct (described later). And in this case, information of a ball is described with a player's touch on a ball (kicking or receiving a ball) as a unit, regarding a touch between different players as a pass and a consecutive touches by the same player as a dribble.

If Fig. 1 is seen further in detail, a region based on color, texture and movement is cut from video sequence 101 and 102 (131). While this process is basically automated, it is also possible to modify it in each field by a tool for correcting regions extracted erroneously and oversplit regions.

Hereafter, the split moving regions are handled as video objects. Also, an object ID can be inserted at this time. Tracing of a region between neighboring fields is automatically determined by a size of the region where split regions are overlapping. A ball, as a special object, is

not currently cut, and manual input is performed with a tool for inputting a position of a ball on an image. 111 in Fig. 1 shows a conceptual diagram of the above data.

Next, a camera movement parameter is restored by tracing in each field a corner or a feature region on a still object in the background.

Position of a video object is represented on an image plane, which requires correction according to movement of a camera. Therefore, a virtual plane is assumed and the video of which a movement parameter is restored is projected on an image plane. Thus, data equivalent to restoring an object position in video which is input from a single camera is acquired. Moreover, since the camera is set to have an angle of depression against the ground, it can be handled not as a distance on an image plane but as a real distance by projecting the object position on the plane of the ground looked down on from the sky as reference plane. This is represented in Space 123 in Fig. 1 as a conceptual diagram which is continuous with time.

While the cut object is a video changing shape over time, in

this example, a silhouette disregarding internal color information is used in order to pay attention to changes in shape. Since changes in silhouette show specific changes according to actions, consecutive silhouettes of plural pre-instructed action patterns are unfolded in unique space so as to acquire changes unique to actions from high order eigenvalue.

Next, when input patterns are unfolded in the same unique space, the movement patterns are identified by seeking which instructed patterns they are closest to. This process requires a changing point in actions in a series of object actions to be acquired. While it is possible to acquire the changing point from the unique space, it is currently entered manually.

A series of objects can be described by plural actions, and just by entering an action ID and a frame number at a changing point in actions, it can be interpreted that all the objects in between were performing the same movements. As above, while the data transformation described in Fig. 1 has processing which partly needs manual input support, it

is possible to generate data in a process which is expected to be automated in the future.

Fig. 2 is a diagram showing in detail the same drawing as TIME (124) in Fig. 1, but the subject video data is different.

Horizontal axis 292 in Fig. 2 represents time, and vertical axis 291 represents an object ID.

An object is described making the changing points of time of silhouettes 201 and 202 in Fig. 2 boundaries of description with an identified action ID as a minimum unit. The time intervals and object positions at the start/end of an action ID are described. Also, in order to trace an object trajectory, positions at plural points in a time interval are described. Accordingly, object positions can be approximated in all the frames.

200 in Fig. 2 is a conceptual diagram of objects according to the time intervals where Objs. 1-6 (203-208) represent the players, (A) and (B) represent the teams, and Obj. X ~~(209) represents the ball~~.

Ball 209 is described as an object without an action ID.

For instance, Obj. 2 (A) is a player of team A and it runs during time interval 214, stays in time interval 224, kicks a ball in time interval 234, runs in time interval 244, kicks a ball in time interval 254, and runs in time interval 264. These actions are recorded with starting time, ending time and an object position (Px).

Fig. 3 shows movement of major players in the soccer scene for about 20 seconds in a solid line and movement of a ball in a dotted line. While Fig. 3 is a diagram showing in detail the same drawing as Space (123) in Fig. 1, but the subject video data is different.

These are the cases where the objects were extracted from the actual soccer scene and then the movement of each object on the field was reproduced by automatically extracting movement parameters of the camera. In the case of Fig. 3, a soccer field is set as reference plane.

The ball object (dotted line) is kicked first from position 312 and reaches position 314 near goal 340 after being kicked several times.

It shows that a player (solid line) has moved from position

322 to position 324 and another player has started from position 312 and moved to position 314.

In this way, a trajectory of each object in the video data on reference plane can be traced.

Next, while in general, plural objects simultaneously exist in an image and have different lifetimes and meanings respectively, it becomes possible to give meaning to a scene consisting of behavior of plural objects.

This is described as Interactive Action (IAction). IAction (or described as IAct) is totally dependent on contents, and a different definition for each content or for each video database administrator may be used.

However, in order to facilitate consistency of description of IAction and its application to a search engine, it is prescribed in this specification that IAction is defined by logical operation with another IAction and plural Actions.

Description of IAction is shown in attached Table 3.

An example of description using this IAction is as follows.

IAct("Pass", 20(frame), 35(frame), 2, 1, 2, Trajectory, 2,

(120, 180, 0) (160, 230, 15))

The above description of IAction is a description of "when, where, who did what." The differences from the above Action are that its subject is plural objects and that the location is specified in a voluntary manner.

In the example of soccer mentioned later, a trajectory of a ball is used as spatial representation of an event.

As an example, descriptions of a "Pass" to which meaning is given by two objects and a "Through Pass" represented by AND of other objects are explained.

Here, it is assumed that the database described by the above Act and IAct actually exists. It is defined for convenience' sake that a through pass is "a pass made between players A and B, and the ball went through between players C and D of the defense side who were there at the time the pass was made."

In the present invention, it is defined as to determine the following. 1. There was an IAct ("pass") between certain two players. 2. There were two other players of the defense

side in that time interval (there are two Acts sharing a time interval with 1, and the object which performed the two Acts belongs to a different team from the objects of 1). 3. The trajectory of the pass went through between the two players (in the shared time interval, the trajectory of the ball intersects the line linking the two players of the defense side).

Here, the IAct statement is described in attached Table 4.

An example of definition of an IAct through pass is shown in attached Table 5.

As an example of application of the present invention, description in a soccer game is shown. In soccer, a type of play defined from a single object and a type of actions defined by relationship of plural objects are used as description items. Moreover, in soccer, description of a ball object is used as a special object.

Description of a ball itself can be represented likewise, namely an action ID is omitted in the Action description of the above object and the object ID is a ball.

The following description defines IAction in soccer

contents, and description using Action and IAction is also possible to other contents.

It can be considered a useful description method for the contents such as the present invention in which representation of behavior of an object and positional relationship on reference plane is effective, since there are a number of examples of application such as sports other than soccer, measurement of traffic and video surveillance, etc.

Here, a ball object is described as follows.

ball (StartTime, EndTime, number of Nodes, Trajectory)

There are the following types of actions of a single object used in soccer.

Activity = { lie, sit, fall_down, raise_hand, dive, hand_throw, throw_in, jump, stay, walk, run, sliding, kick, overhead_kick }

Namely, lie, sit, fall_down, raise_hand, dive, hand_throw, throw_in, jump, stay, walk, run, sliding, kick, overhead_kick, and so on can be considered in general.

As examples of actions, a pass (lines 1 to 26 in Table 6), a long pass (lines 28 to 39 in Table 6), a feed pass (lines 1 to 14 in Table 7), a cross pass (lines 16 to 33 in Table 7), a gain pass (lines 1 to 18 in Table 8), centering (lines 20 to 38 in Table 8), and a wall pass (lines 1 to 36 in Table 9) are described.

As definition of a through pass as an example of IAct was already explained in detail referring to Table 5, detailed explanation of these examples is not given here.

In addition, the groups of auxiliary functions used in IAct are shown in attached Tables 10 and 11. See the groups of auxiliary functions for the meanings of the functions used in the above examples (see attached Tables 6 to 9).

Fig. 4 shows the steps necessary for searching video data and related data groups (see attached Tables 10 and 11).

Video data 401 is processed by image processing 410 and the results are stored as data groups of 420. For instance,

region map 422 is acquired from a region cutting process, an object trajectory ID (424) from identification of an object, an action ID (426) from classification of actions, camera parameter 428 from camera action restoring process, etc. However, these processes are not only completely automated processes but also the cases of formatting data with manual support after generation and manually inputting as data directly from video data 401.

From these data groups of 420, Act 430, namely video description according to the objects is obtained.

Also, from the definition obtained in advance by selecting video data 401, description of reference plane (Refplane) 442 for video data 401 is obtained.

Furthermore, by selecting video data 401, scene description IAct 456 describing meaning given to a scene consisting of behavior of plural objects is obtained, which is predefined for interpreting video data 401.

As an application of this description, to search video data 401, a user will enter user keyword 470 to search engine 460. Search engine 460 interprets user keyword 470 and

returns a time interval of corresponding video data 401 from object description 430 and scene description 456 to display on video to the user.

At this time, scene description 456 returns a corresponding time interval by processing Refplane 422 and Act 430.

Moreover, a user is allowed a description equivalent to scene description IAct 456 for search engine 460, and Refplane 422 and Act 430 are processed for a user-defined scene description so that a time interval of a user-defined scene is returned and displayed on video to the user.

Fig. 5 shows a screen for searching video.

It is possible for a user to search a desired scene by selecting a necessary item in search screen 500 and starting search.

In the case of a soccer video explained so far, it is possible to perform specification of players 510, specification of time 520, specification of location (place) 530, and specification of action 540.

As to specification of players, it is thinkable to specify them by team name, individual name or position.

In the case of specification of action, it is possible to use the actions defined in the above-mentioned Action and IAction. For instance, it is possible to use Actions such as lie, sit, fall_down, raise_hand, dive, hand_throw, throw_in, jump, stay, walk, run, sliding, kick or overhead_kick, or IActions such as a pass, a through pass or a centering. Moreover, a user may also newly define a scene.

Fig. 6 shows screen 600 of a search result of a scene in the case of specifying a through pass for an action in Fig. 5. For instance, the search result in this case is one, and image 610 at the start of the scene is displayed. Generally, a desired scene is replayed by clicking this image 610 of the search result.

Advantages of the Invention

It is possible, by adopting the above-mentioned organization of the present invention, to provide a description method for effectively representing contents of motion picture with a small data volume.

Also, a description method effective for interpretation based on contents of motion picture becomes possible by adopting the organization of the present invention.

Furthermore, it becomes possible, by adopting the organization of the present invention, to provide a description method of motion picture capable of applications such as reuse of an object and summarization of contents in addition to search for an object or a scene.

As a summary, other embodiments are described below.

A description means for contents of motion picture, the means comprising of:

(a) means for setting reference plane;

(b) means for describing each object on the motion picture by position on the reference plane and predefined type of actions; and

(c) means for describing each scene by using the means for describing each object.

A search means for contents of motion picture, the means comprising of:

(a) means for setting reference plane;

(b) means for describing each object on the motion picture by position on the reference plane and predefined type of actions;

(c) means for describing each scene by using the means for describing each object; and

(d) means for searching motion picture by using the means for describing each object or the means for describing each scene.

A description method for motion picture, the method comprising the steps of:

(a) determining reference plane which represents information of object positions included in the motion picture;

(b) representing changes over time of each object on the reference plane as a trajectory;

(c) setting a description unit based on predefined type of actions of each object by using changes in shape of each

object so as to assign actions of each object as each behavioral section; and

(d) defining each scene by plural objects.

A search method for motion picture, the method comprising the steps of:

(a) setting reference plane which represents information of object positions included in the motion picture;

(b) representing changes over time of each object on the reference plane as a trajectory;

(c) setting a description unit based on predefined type of actions of each object by using changes in shape of each object so as to assign actions of each object as each behavioral section;

(d) defining each scene by plural objects; and

(e) searching a specific scene by using the actions of each object or the scene.

A description method for motion picture, the method comprising the steps of:

(a) determining reference plane from the motion picture;

(b) cutting a region map, an object trajectory ID, an action ID and a camera parameter from the motion picture;

(c) creating description of actions by each object from the region map, the object trajectory ID, the action ID and the camera parameter; and

(d) creating description of scenes by using the description of actions by each object.

A description method for motion picture, the method comprising the steps of:

(a) determining reference plane from the motion picture;

(b) cutting a region map, an object trajectory ID, an action ID and a camera parameter from the motion picture;

(c) creating description of actions by each object from the region map, the object trajectory ID, the action ID and the camera parameter; and

(d) creating description of scenes by using the description of actions by each object.

A description method for motion picture, the method comprising the steps of:

(a) cutting a region map, an object trajectory ID, an action ID and a camera parameter from the motion picture;

(b) creating description of actions by each object from the region map, the object trajectory ID, the action ID and the camera parameter; and

(c) creating description of scenes by using the description of actions by each object.

A computer readable storage medium which has recorded management data for searching motion picture, the management data comprising of:

(a) data of description of actions by each object defined by position on reference plane and predefined type of actions; and

(b) data of description of scenes defined by the data of description of actions.

A computer readable storage medium which has recorded a

program, the program having a computer execute the steps of:

(a) determining reference plane which represents information of object positions included in the motion picture;

(b) representing changes over time of each object on the reference plane as a trajectory;

(c) setting a description unit based on predefined type of actions of each object by using changes in shape of each object so as to assign actions of each object as each behavioral section; and

(d) defining each scene by plural objects.

Description of the Symbols

401: Video data

410: Image processing

420: Data groups

422: Region map

424: Object trajectory ID

426: Action ID

428: Camera parameter

430: Act description

442: Reference plane description

456: IAct description

460: Search engine

470: User keyword